

# Regressione Lineare con un Singolo Regressore

- Quali sono gli effetti dell'introduzione di pene severe per gli automobilisti ubriachi?
- Quali sono gli effetti della riduzione della dimensione delle classi sui punteggi degli studenti in un test?
- Quali sono gli effetti dell'istruzione sul reddito futuro?
- Tutte queste domande fanno riferimento al modello di regressione lineare.

- Tale modello ipotizza l'esistenza di una relazione lineare tra due variabili  $X$  ed  $Y$ .
- La pendenza della retta che mette in relazione tali variabili fornisce l'effetto di una variazione unitaria di  $X$  su  $Y$ .
- Il nostro compito sarà quello di stimare tale pendenza e, quindi, di valutare l'effetto su  $Y$  di una variazione unitaria di  $X$  basandoci su di un campione di dati relativi a tali variabili.
- La pendenza della retta e la sua intercetta saranno stimate utilizzando il metodo dei Minimi Quadrati Ordinari (OLS).

## Il Modello

- Supponiamo di essere i consulenti di un direttore scolastico il quale deve decidere se assumere altri insegnanti.
- Attuando tale politica ridurrebbe di due unità il rapporto studenti/insegnanti
- Da un lato i genitori vogliono classi più piccole per garantire una maggiore attenzione personale ai loro figli
- Dall'altro lato assumere altri insegnanti comporta un sensibile aumento dei costi
- Il nostro compito è rispondere alla seguente domanda:
  - Se si taglia la dimensione delle classi, quale sarà l'effetto sul rendimento degli studenti?

## Il Modello

- Supponiamo che il rendimento degli studenti sia misurato tramite i punteggi ottenuti in un test standardizzato. La domanda diventa
  - Se si taglia la dimensione, quale sarà l'effetto sui punteggi del test nel distretto d'appartenenza?
- Rappresentiamo matematicamente il problema del provveditore

$$\beta_{ClassSize} = \frac{\Delta TestScore}{\Delta ClassSize}$$

- Se conoscessimo  $\beta_{ClassSize}$  potremmo immediatamente rispondere alla domanda posta dal provveditore.
- Reimpostando il precedente rapporto abbiamo

$$\Delta TestScore = \beta_{ClassSize} \times \Delta ClassSize$$

## Il Modello

- Se  $\beta$  è uguale a -0.6, l'effetto di una riduzione della dimensione pari a due unità comporta un aumento del punteggio pari a  $(-0.6) \times (-2) = 1.2$
- Possiamo ora definire la retta che mette in relazione i risultati del test e la dimensione delle classi.

$$TestScore = \beta_0 + \beta_{ClassSize} * ClassSize$$

- Tuttavia, il provveditore potrebbe non essere soddisfatto dalla vostra risposta. Perché?

## Il Modello

- La dimensione della classe è soltanto uno dei molteplici aspetti che possono influenzare l'istruzione elementare.
- Due distretti con classi della stessa dimensione possono avere punteggi diversi perché ha insegnati migliori, testi migliori, studenti immigrati, etc...
- Infatti la relazione da noi evidenziata non può valere per ogni distretto ma rappresenta una relazione valida in media nella popolazione dei distretti.

## Il Modello

- Una relazione che valga per ciascun distretto deve incorporare tutti gli altri fattori che possono influenzare i punteggi dei test.

$$TestScore = \beta_{ClassSize} * \Delta ClassSize + altrifattori$$

- Introduciamo una versione più generale

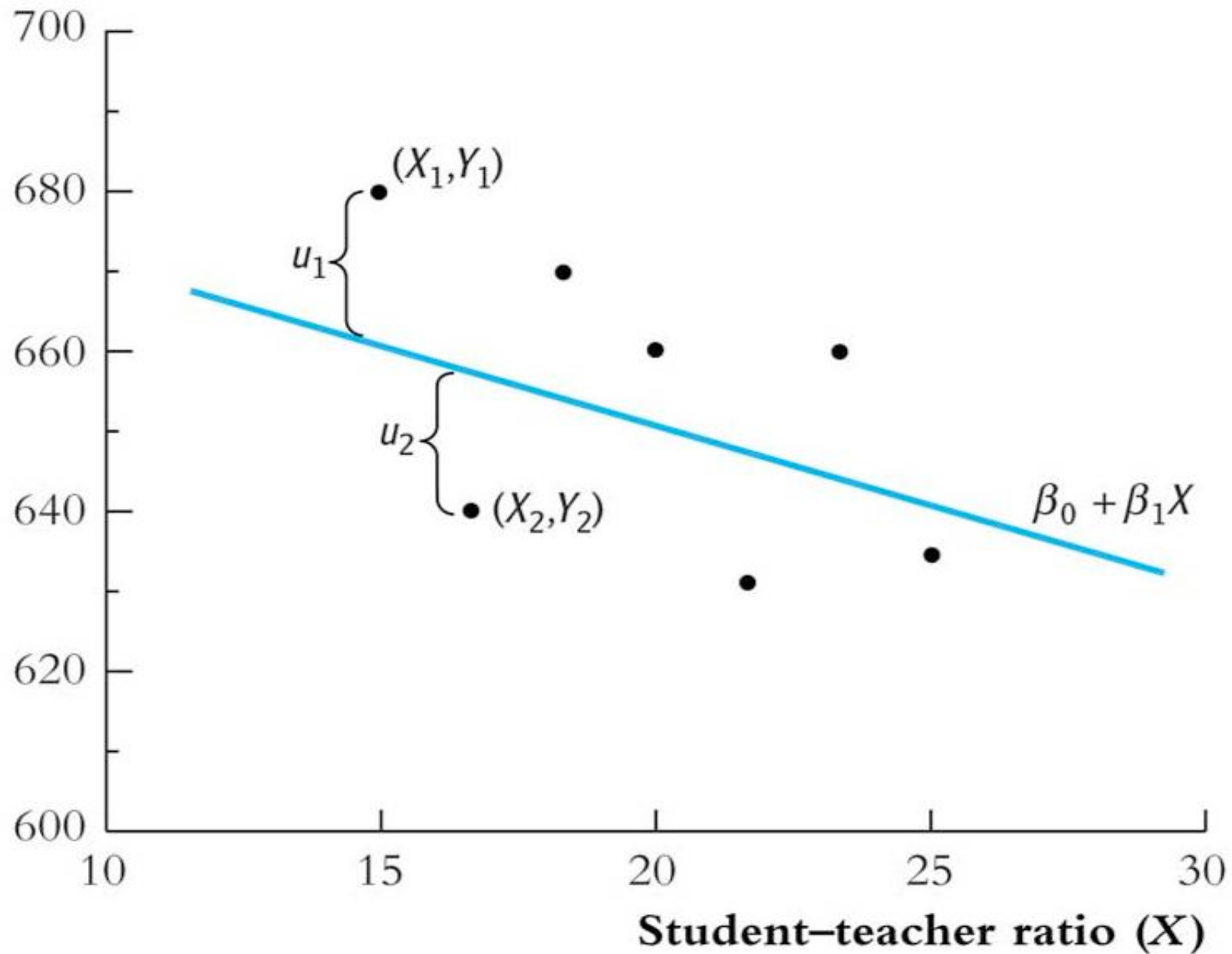
$Y_i = \beta_0 + \beta_1 X_i + u_i$  per  $i = 1, \dots, n$  dove  $\beta_0$  è l'intercetta e  $\beta_1$  è la pendenza

- Tale equazione è il modello di regressione lineare con singolo regressore, in cui  $Y$  è la variabile dipendente ed  $X$  è la variabile indipendente o regressore.
- I primi due componenti del membro di destra costituiscono la retta di regressione della popolazione.

## Il Modello

- L'intercetta  $\beta_0$  e la pendenza  $\beta_1$  sono i coefficienti o parametri della retta di regressione della popolazione
- L'intercetta è il valore della retta di regressione quando  $X=0$ . In alcune applicazioni economiche, l'intercetta ha un significato concreto mentre in altre ne è priva.
- Il termine  $u_i$  è l'errore o disturbo, ed incorpora tutti i fattori responsabili della differenza tra il punteggio medio del test dell' $i$ -esimo distretto ed il valore predetto dalla retta di regressione.

**Test score (Y)**



- La figura mostra il modello di regressione lineare con un singolo regressore considerando sette ipotetiche osservazioni sul punteggio del test (Y) e la dimensione delle classi (X).
- La retta di regressione della popolazione,  $\beta_0 + \beta_1 X$ , ha pendenza negativa ( $\beta_1 < 0$ ), mostrando che distretti con rapporto studenti/insegnati minore tendono ad avere punteggi maggiori.
- L'intercetta in questo caso non ha alcun significato concreto
- A causa degli altri fattori che influenzano i risultati del test, le osservazioni ipotetiche non cadono esattamente sulla retta di regressione della popolazione.

- Tornando al nostro compito di consulenti, qual è l'effetto atteso sui punteggi del test di una riduzione di 2 unità nel numero di studenti per insegnanti?
  - La variazione attesa è  $(-2) * \beta_{\text{ClassSize}}$
- Qual è il valore di  $\beta_{\text{ClassSize}}$ ?

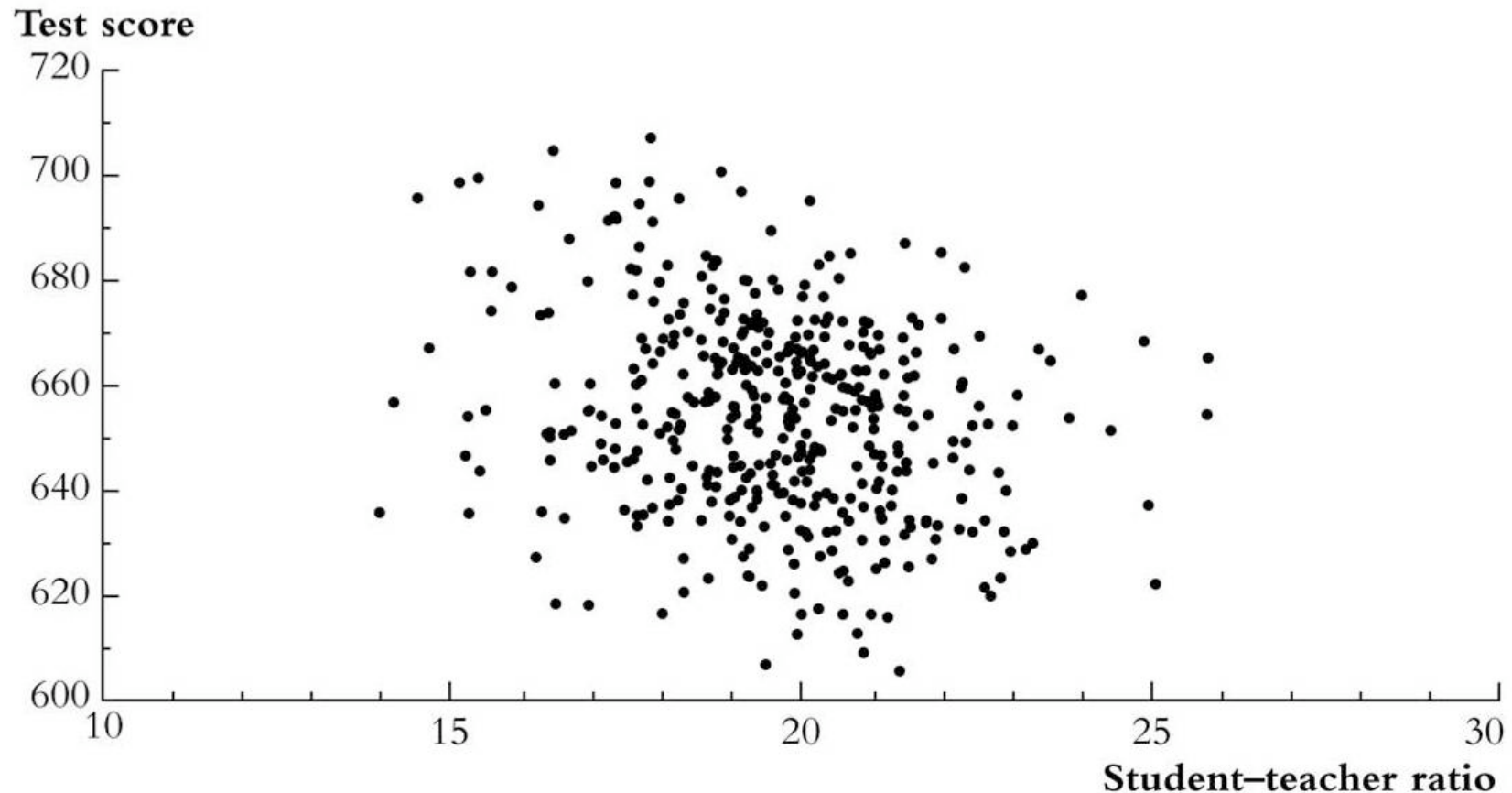
# Stima dei coefficienti

- Come in ogni situazione pratica i coefficienti della retta di regressione sono ignoti.
- Possiamo studiare la relazione tra Y ed X usando un campione di dati relativi alle due variabili.

|                  |       |              | Percentile |      |       |                |       |       |       |
|------------------|-------|--------------|------------|------|-------|----------------|-------|-------|-------|
|                  | Media | Dev.<br>Std. | 10%        | 25%  | 40%   | 50%<br>Mediana | 60%   | 75%   | 90%   |
| Stud/in<br>segn. | 19.6  | 1.9          | 17.3       | 18.6 | 19.3  | 19.7           | 20.1  | 20.9  | 21.9  |
| Punt.<br>test    | 654.2 | 19.1         | 630.4      | 640  | 649.1 | 654.5          | 659.4 | 666.7 | 679.4 |

# Stima dei coefficienti

- La correlazione campionaria ( $-0.23$ ) indica una debole relazione negativa tra le variabili.
- Sembrano esserci altre determinanti dei punteggi che impediscono alle osservazioni di disporsi lungo una linea retta



- Potremmo in ogni caso provare a tracciare una retta attraverso la nuvola di punti, ma ognuno potrebbe tracciare una linea diversa.
- Come possiamo scegliere tra le diverse linee rette possibili?
  - Lo stimatore dei minimi quadrati ordinari (OLS)

# Lo Stimatore dei Minimi Quadrati Ordinari (OLS)

- Tale stimatore sceglie i coefficienti di regressione in modo che la retta di regressione stimata sia il più possibile vicina ai dati osservati, dove la vicinanza è misurata dal quadrato degli errori che si commettono nel predire  $Y$  data  $X$ .
- Lo stimatore estende l'idea esposta nel caso della media campionaria al modello di regressione lineare.

- Siano  $b_0$  e  $b_1$  gli stimatori di  $\beta_0$  e  $\beta_1$ . La retta di regressione basata su tali stimatori è  $b_0 + b_1 X$
- Il valore di  $Y_i$  predetto usando questa retta è  $b_0 + b_1 X_i$ .
- L'errore che si commette nel predire  $Y_i$  è  $Y_i - (b_0 + b_1 X_i) = Y_i - b_0 - b_1 X_i$
- La somma dei quadrati degli errori di predizione per tutte le  $n$  osservazioni è

$$\sum_{i=1}^n (Y_i - b_0 - b_1 X_i)^2$$

- Gli stimatori dell'intercetta e della pendenza che minimizzano la somma dei quadrati degli errori sono detti stimatori dei minimi quadrati ordinari di  $\beta_0$  e  $\beta_1$ .
- Lo stimatore OLS di  $\beta_0$  e  $\beta_1$  sono indicati con  $\hat{\beta}_0$  e  $\hat{\beta}_1$
- La retta di regressione degli OLS è  $\hat{\beta}_0 + \hat{\beta}_1 X$
- Il valore predetto di  $Y_i$  data  $X_i$  è  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$
- Il residuo relativo alla  $i$ -esima osservazione è la differenza tra  $Y_i$  ed il suo valore predetto,  $\hat{u}_i = Y_i - \hat{Y}_i$

- Formule chiave

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{s_{XY}}{s_X^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

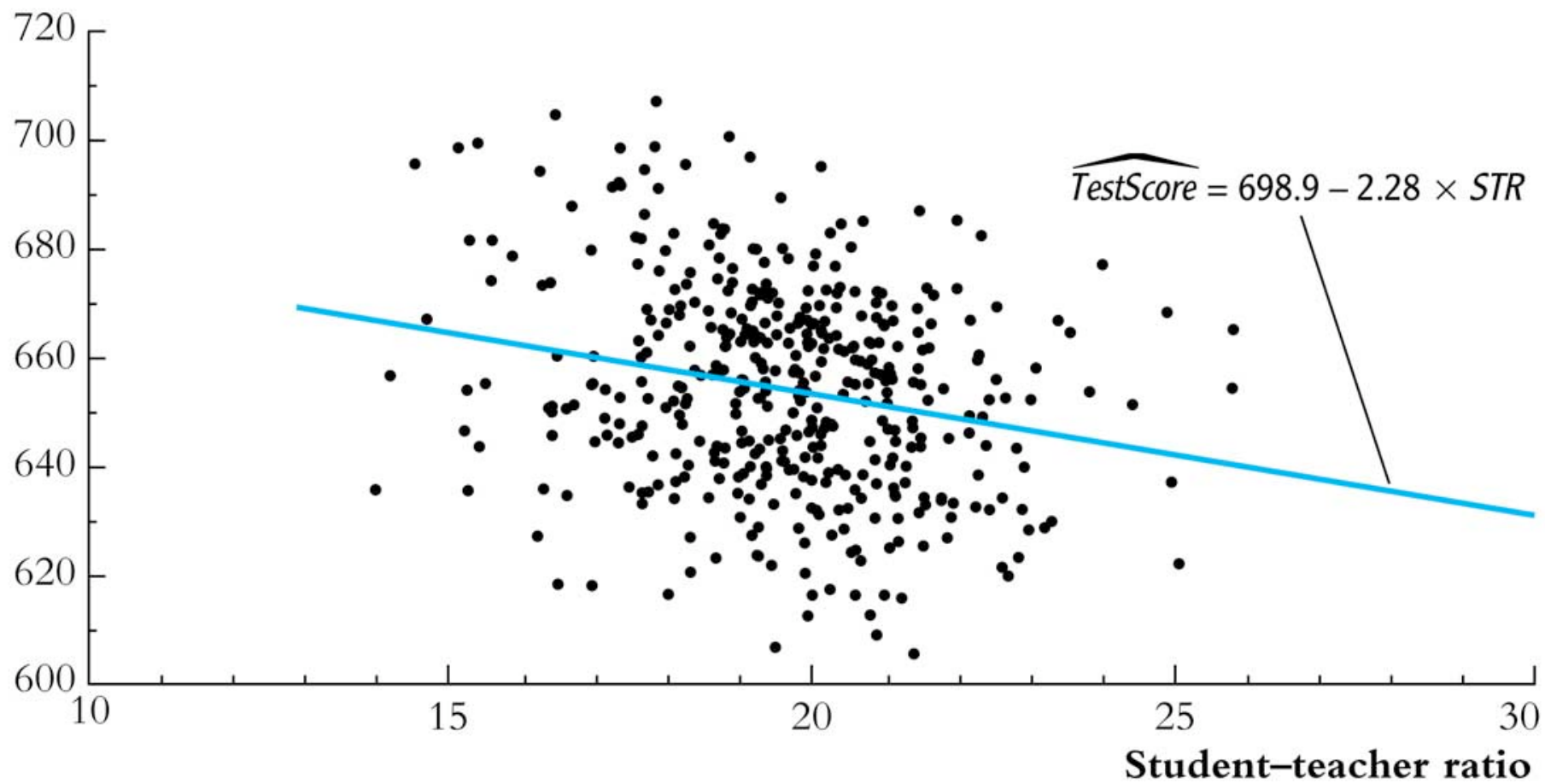
## Esempio di analisi OLS

- Utilizziamo i dati circa il rapporto studenti/insegnanti con i punteggi del test già illustrati
- Applicando gli OLS otteniamo che la pendenza stimata è -2.28 e che l'intercetta è 698.9
- La retta di regressione OLS è

$$\hat{TestScore} = 698.9 - 2.28 \times STR$$

- TestScore è il punteggio medio del test ottenuto nel distretto; STR sta per rapporto studenti/insegnanti
- Il simbolo su TestScore indica che è il valore predetto in base alla regressione degli OLS.

**Test score**



- La pendenza della retta indica che un incremento di uno studente del STR è associato ad una riduzione nei punteggi del test a livello di distretto di 2.28 punti.
- In maniera analoga una riduzione di due studenti è associata ad una riduzione nei punteggi del test di  $(-2 * (-2.28)) = 4.56$
- A questo punto siamo in grado di predire il risultato del test per il distretto dato un determinato valore del STR.
  - Ad esempio con 20 studenti per insegnante il punteggio predetto è  $698.9 - (2.28 * 20) = 653.3$

- Tale predizione non può essere perfetta poiché vi sono tanti altri fattori che influenzano il punteggio del test e che non sono inclusi nella regressione
- Il provveditore sta aspettando la nostra consulenza circa la convenienza di ridurre il STR di due unità assumendo nuovi insegnanti.
- Supponiamo di lavorare per il distretto mediano, quello della California. Dalla [tabella](#) precedente sappiamo che il STR mediano è 19.7 e che il punteggio mediano è 654.5.
- Una riduzione di due unità del STR lo sposterebbe dal 50-esimo al 10-mo percentile.
- Quale potrebbe essere l'effetto sui punteggi?

- La retta di regressione mostra che un taglio del numero di studenti per insegnanti comporta un aumento del punteggio di 4.6 punti.
- Nel caso del distretto mediano si passerebbe da 654.5 a 659.1. È un buon miglioramento?
- Secondo la tabella il distretto si collocherebbe appena sotto il 60-esimo percentile.
- Tagliare il STR in maniera rilevante aiuterebbe e potrebbe essere una buona idea avendo le risorse per assumere nuovi insegnanti.

# Esempio - CAPM

- Abbiamo bisogno di un incentivo finanziario per assumere un rischio
- Il rendimento atteso di un investimento rischioso deve essere maggiore di quello di un investimento sicuro
- $(R - R_f) = \beta(R_m - R_f)$
- $\beta < 1$  implica una bassa redditività del titolo
- $\beta > 1$  indica un maggiore tasso di rendimento atteso rispetto al portafoglio di mercato
- Il  $\beta$  di ogni titolo viene stimato tramite regressioni OLS dell'effettivo eccesso di rendimento del titolo sull'effettivo eccesso di rendimento di un ampio indice di mercato.

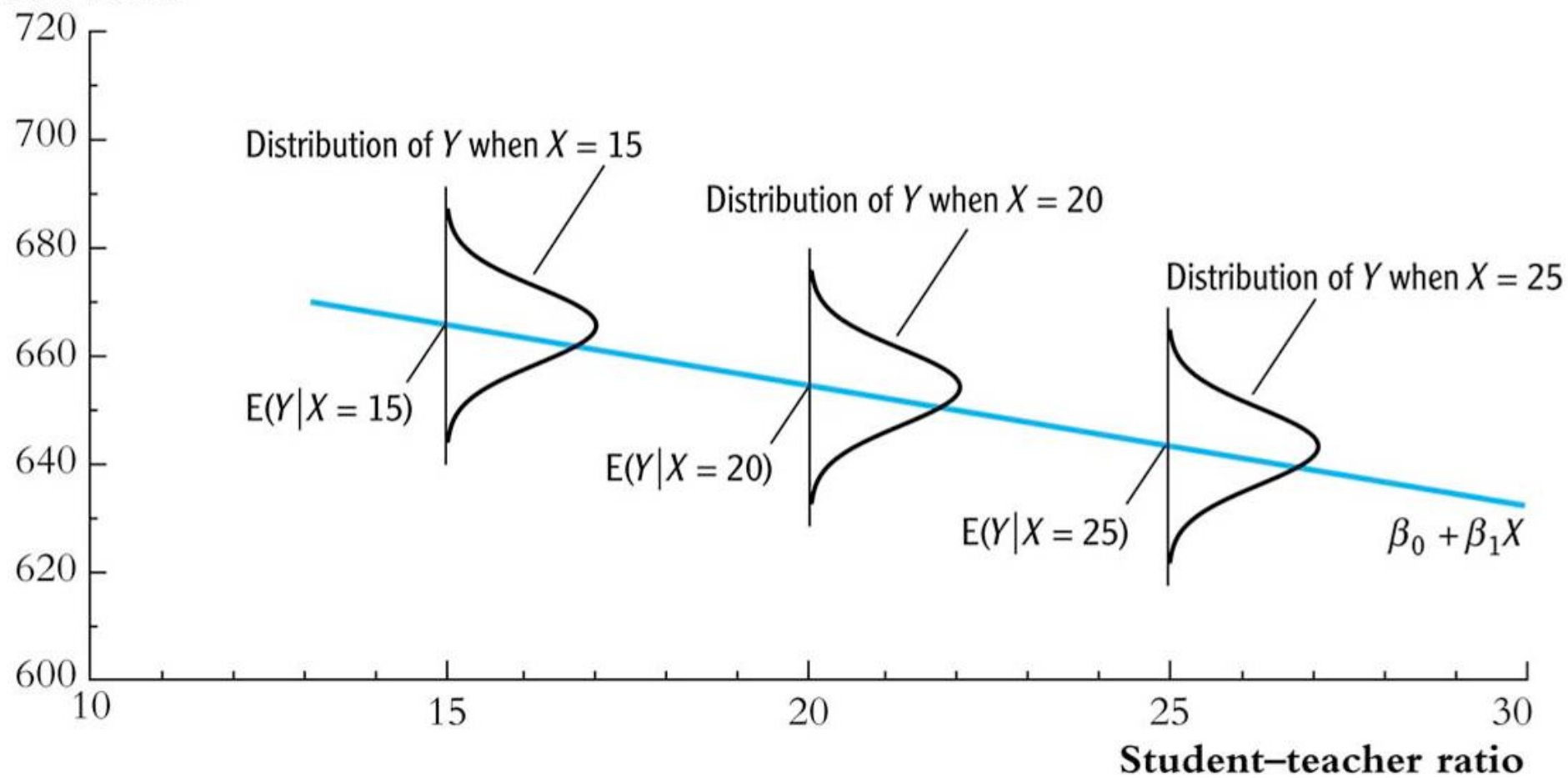
# Assunzioni dei Minimi Quadrati

1. La distribuzione condizionata di  $u_i$  data  $X_i$  ha media nulla
2.  $(X_i, Y_i)$ ,  $i=1, \dots, n$  sono indipendentemente e identicamente distribuite
3.  $X_i$  e  $u_i$  hanno quattro momenti

# Assunzione 1

- Tale assunzione si basa sul fatto che “gli altri fattori” contenuti nel termine di errore  $u_i$  non sono legati a  $X_i$  e, quindi, la media condizionata di tali fattori è zero.
- Dato che la media condizionata è nulla, anche la covarianza sarà nulla e le due variabili casuali non sono correlate.
  - $E(u_i|X_i)=0$  implica  $\text{corr}(X_i, u_i)=0$
- Per una data dimensione delle classi, ad esempio 20 studenti per classe, talvolta tali fattori comportano prestazioni migliori rispetto quelle predette ( $u_i > 0$ ) e talvolta peggiori ( $u_i < 0$ ) anche se la predizione in media è corretta.

**Test score**

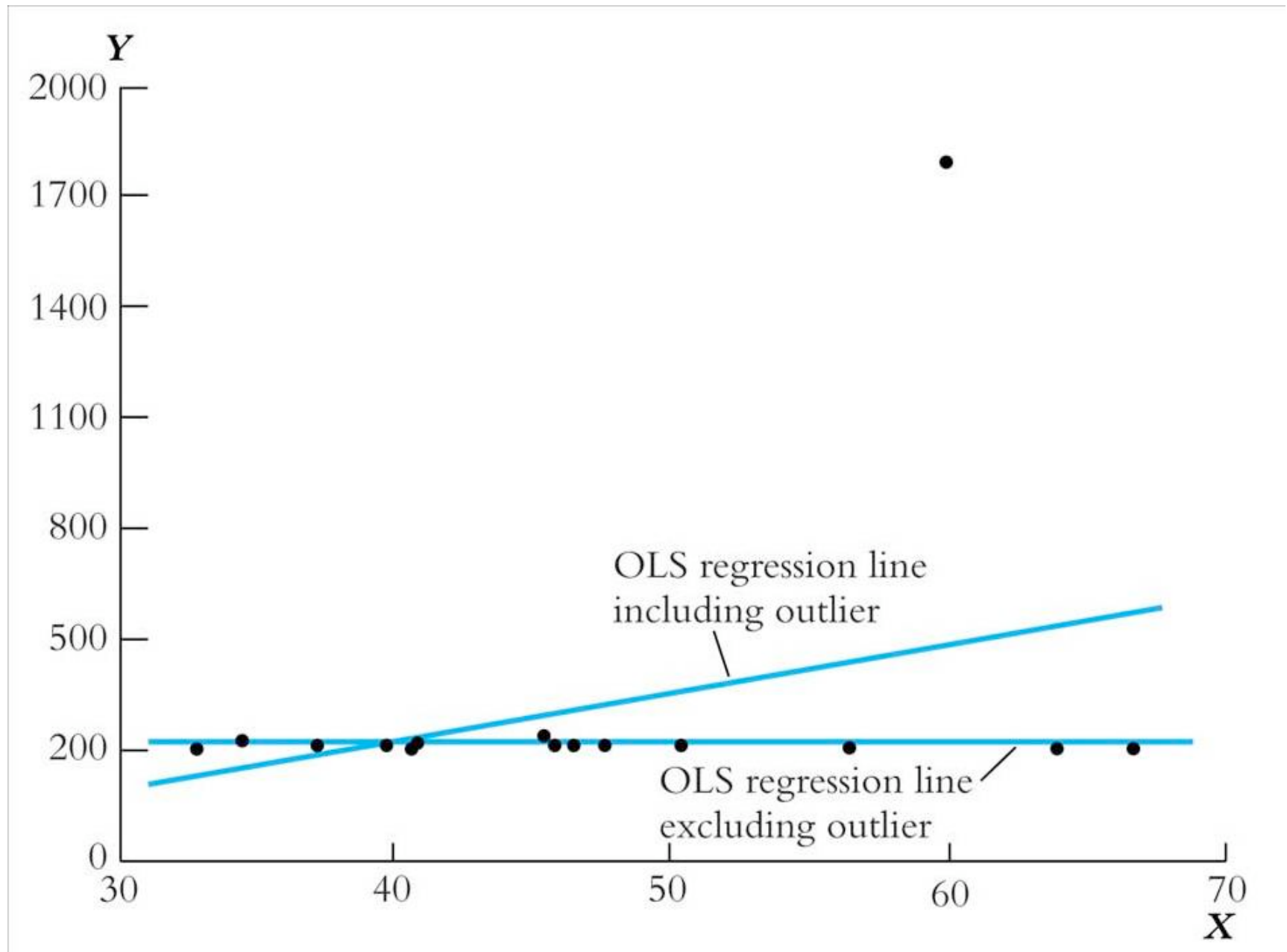


## Assunzione 2

- Come abbiamo già visto tale assunzione deriva dalla modalità di campionamento.
- Se le osservazioni sono estratte tramite campionamento casuale semplice da un'unica popolazione allora  $(X_i, Y_i)$  sono i.i.d
- Diversi sono i casi in cui il valore delle  $X$  viene fissato dal ricercatore per specifici studi.
- Quando le osservazioni si riferiscono alla stessa unità osservata ripetutamente nel tempo.

## Assunzione 3

- $X_i$  e  $u_i$  hanno quattro momenti non nulli e finiti.
- In tale modo si limita la probabilità di selezionare un'osservazione con valori estremamente elevati di  $X_i$  e  $u_i$
- Questa eventualità potrebbe rendere i valori della regressione fuorvianti.



# Distribuzione Campionaria degli stimatori degli OLS

- La media campionaria è uno stimatore non distorto di  $\mu_Y$  e per il teorema del limite centrale la sua distribuzione approssima la normale
- Lo stesso ragionamento può essere applicato ai coefficienti di regressione  $\hat{\beta}_0$  e  $\hat{\beta}_1$
- Tali stimatori sono non distorti e consistenti

## Verifica di Ipotesi su un singolo coefficiente di regressione

- Il solito provveditore vi chiama per chiedervi se vi è qualche evidenza nel vostro campione di 420 osservazioni circa i distretti scolastici californiani che la pendenza della retta di regressione tra dimensione della classe e punteggio del test sia nulla.
- Dobbiamo verifica se la pendenza  $\beta_1$  o l'intercetta  $\beta_0$  è nulla

## Ipotesi bilaterali su $\beta_1$

- $H_0: \beta_1 = \beta_{1,0}$  contro  $H_1: \beta_1 \neq \beta_{1,0}$
- A questo punto operiamo come già fatto
- Calcoliamo l'errore standard di  $\hat{\beta}_1$ ,  $SE(\hat{\beta}_1)$

$$SE(\hat{\beta}_1) = \sqrt{\sigma_{\hat{\beta}_1}^2}$$

- Solitamente tale valore è fornito dai software econometrici

- Poi calcoliamo la statistica  $t$ ,

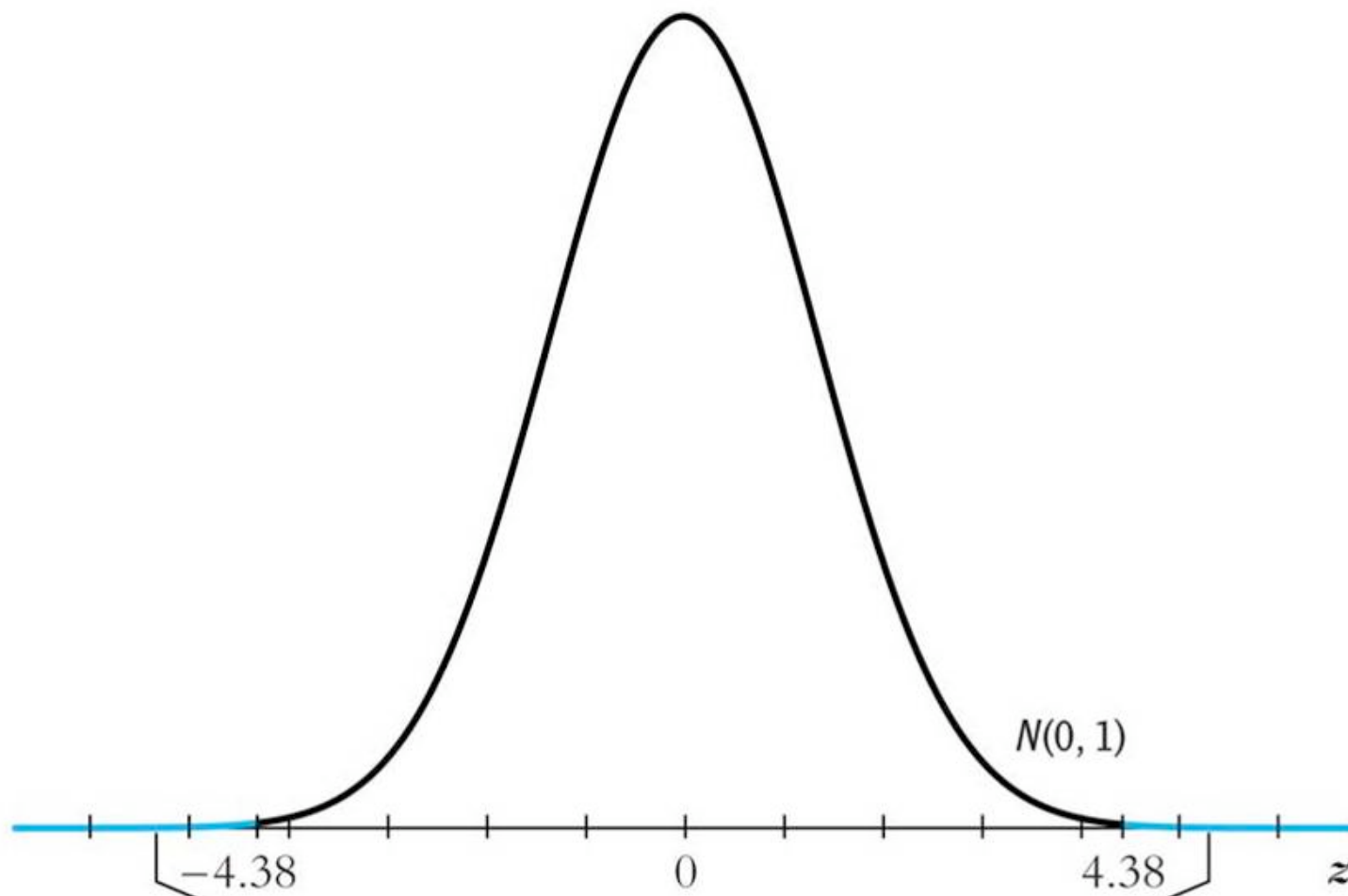
$$t = \frac{\hat{\beta}_1 - \hat{\beta}_{1,0}}{SE(\hat{\beta}_1)}$$

- Infine determiniamo il valore-p, ovvero la probabilità di osservare un valore di  $\hat{\beta}_1$  diverso da  $\hat{\beta}_{1,0}$  almeno quanto la stima realmente ottenuta, assumendo che l'ipotesi nulla sia corretta.
- Valore-p =  $\Pr(|Z| > |t^{\text{act}}|) = 2\Phi(-|t^{\text{act}}|)$
- Allo stesso modo possiamo utilizzare la statistica  $t$  ed i valori critici associati al livello di significatività scelto

- Tornando al nostro esempio sappiamo che la pendenza della retta è -2.28. Il suo errore standard  $SE(\hat{\beta}_1)=0.52$ .
- Volendo testare l'ipotesi nulla che  $\beta_{\text{ClassSize}}=0$ , possiamo calcolare la statistica t
- $t^{\text{act}} = (-2.28-0)/0.52 = -4.38$
- Tale statistica è maggiore del valore critico bilaterale all'1% e quindi rifiutiamo l'ipotesi nulla
- Allo stesso modo, il valore-p associato è estremamente piccolo 0.0001.

# Ipotesi unilaterale

- Supponiamo di voler testare l'ipotesi secondo la quale classi più piccole favoriscano l'apprendimento.
- Sotto tale ipotesi  $\beta_1 < 0$ . Quindi possiamo scrivere
  - $H_0: \beta_1 = \beta_{1,0}$  contro  $H_1: \beta_1 < \beta_{1,0}$
- I passi da seguire sono gli stessi già visti
  - Rifiutiamo  $H_0$  se  $|t^{\text{act}}| < -1.645$  al 5% oppure se il valore-p =  $\Pr(Z < t^{\text{act}}) = \Phi(t^{\text{act}})$
- Nel nostro esempio,  $t^{\text{act}} = -4.38$  che è  $<$  di  $-2.33$  (valore critico unilaterale all'1%) e quindi si rifiuta l'ipotesi nulla al livello dell'1%.



The  $p$ -value is the area  
to the left of  $-4.38$   
+  
the area to the right of  $+4.38$ .

# Intervalli di confidenza per un coefficiente di regressione

- Gli intervalli sono costruiti in maniera analoga a quanto già visto per lo studio di un campione.

$$\hat{\beta}_1 \pm (1.96 \times SE(\hat{\beta}_1))$$

- Applicando la formula al nostro esempio, l'intervallo di confidenza al 95% per  $\beta_1$  è  $(-2.28 \pm 1.96 \times 0.52) = (-3.30, 1.26)$
- Dato che il valore  $\beta_1 = 0$  non è contenuto nell'intervallo di confidenza possiamo rifiutare l'ipotesi nulla.

# Regressione quando $X$ è una variabile binaria

- Una variabile binaria è anche detta variabile indicatrice o variabile dummy.
  - Lavoratore (=1 se femmina, =0 se maschio)
- L'interpretazione dei coefficienti della regressione è, in questo caso, differente.
- Supponiamo di avere una variabile  $D_i$  che è uguale a 0 se il rapporto studenti/insegnanti è  $<20$  oppure a 1 se tale rapporto è  $\geq 20$ .
- Il modello della regressione della popolazione è
- $Y_i = \beta_0 + \beta_1 D_i + u_i$  con  $i=1, \dots, n$

- Dato che il regressore non è continuo non possiamo pensare al coefficiente  $\beta_1$  come alla pendenza di una retta. Verrà quindi indicato semplicemente come il coefficiente di  $D_i$
- Il modo migliore per interpretare tale coefficiente è analizzare uno alla volta i due casi possibili.
- Supponiamo il caso in cui il rapporto studenti-insegnanti è alto ( $D_i=0$ ), allora la retta diventa  $Y_i=\beta_0+u_i$
- La media condizionata di  $Y_i$  quando  $D_i=0$  è  $E(Y_i|D_i=0)=\beta_0$ , cioè  $\beta_0$  è il punteggio medio del test nella popolazione quando il rapporto studenti-insegnanti è alto

- Allo stesso modo quando  $D_i=1$  abbiamo la seguente retta  $Y_i=\beta_0+\beta_1+u_i$
- In questo caso avremo  $E(Y_i|D_i=1)=\beta_0+\beta_1$ , questo sarà il punteggio medio nel caso in cui il rapporto studenti-insegnanti è basso.
- $\beta_1$  è anche la differenza tra la media condizionata di  $Y_i$  quando  $D_i=1$  e quando  $D_i=0$ . Quindi il suo stimatore OLS è la differenza tra la media campionaria di  $Y_i$  tra i due gruppi.

# Esempio

- Nel nostro esempio una regressione con la variabile dummy è data da

$$\hat{TestScore} = 650,0 + 7.4D$$

(1.3)            (1.8)

- Quale è il punteggio medio ottenuto dal sottocampione con rapporto studenti-insegnati alto? Quale è il punteggio medio relativo al caso di rapporto basso?
- Quale è la differenza tra punteggi medi?
- È tale differenza significativa al 5%?

## Esempio

- Se  $D_i=0$  allora il punteggio medio è 650
- Se  $D_i=1$  il punteggio medio è 657.4
- La differenza è 7.4
- La statistica  $t$  nel caso di ipotesi nulla circa differenza pari a zero è  $7.4/1.8=4.04>1.96$   
rifiutiamo l'ipotesi nulla
- Intervallo di confidenza:  $7.4\pm 1.96\times 1.8=(3.9, 10.9)$   
dato che tale intervallo esclude il valore 0,  
possiamo rifiutare l'ipotesi nulla.

## R<sup>2</sup> di regressione

- È una misura della bontà con cui la retta di regressione degli OLS interpola i dati
- È la frazione della varianza campionaria di  $Y_i$  spiegata da  $X_i$ .
- La variabile dipendente può anche essere scritta come somma del valore predetto e del residuo
- $$Y_i = \hat{Y}_i + u_i$$
- L'R<sup>2</sup> può essere scritto come il rapporto tra la somma dei quadrati (varianza) spiegata e quella totale.

- La somma dei quadrati spiegata (ESS) è la somma dei quadrati degli scarti dei valori predetti dalla loro media.

$$ESS = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2$$

- La somma dei quadrati totale (TSS) è la somma degli scarti dei valori osservati dalla loro media

- 

$$TSS = \sum (Y_i - \bar{Y})^2$$

- Sulla base di tali formule l' $R^2 = ESS/TSS$
- Alternativamente, l' $R^2$  può essere scritto utilizzando la frazione di varianza non spiegata da  $X_i$ . Cioè dalla somma dei quadrati dei residui (RSS) degli OLS

$$SSR = \sum_{i=1}^n \hat{u}_i^2$$

- Sapendo che  $TSS = ESS + SSR$ , possiamo scrivere che  $R^2 = 1 - (SSR/TSS)$
- $R^2$  varia da 0 ad 1

# Errore standard della regressione

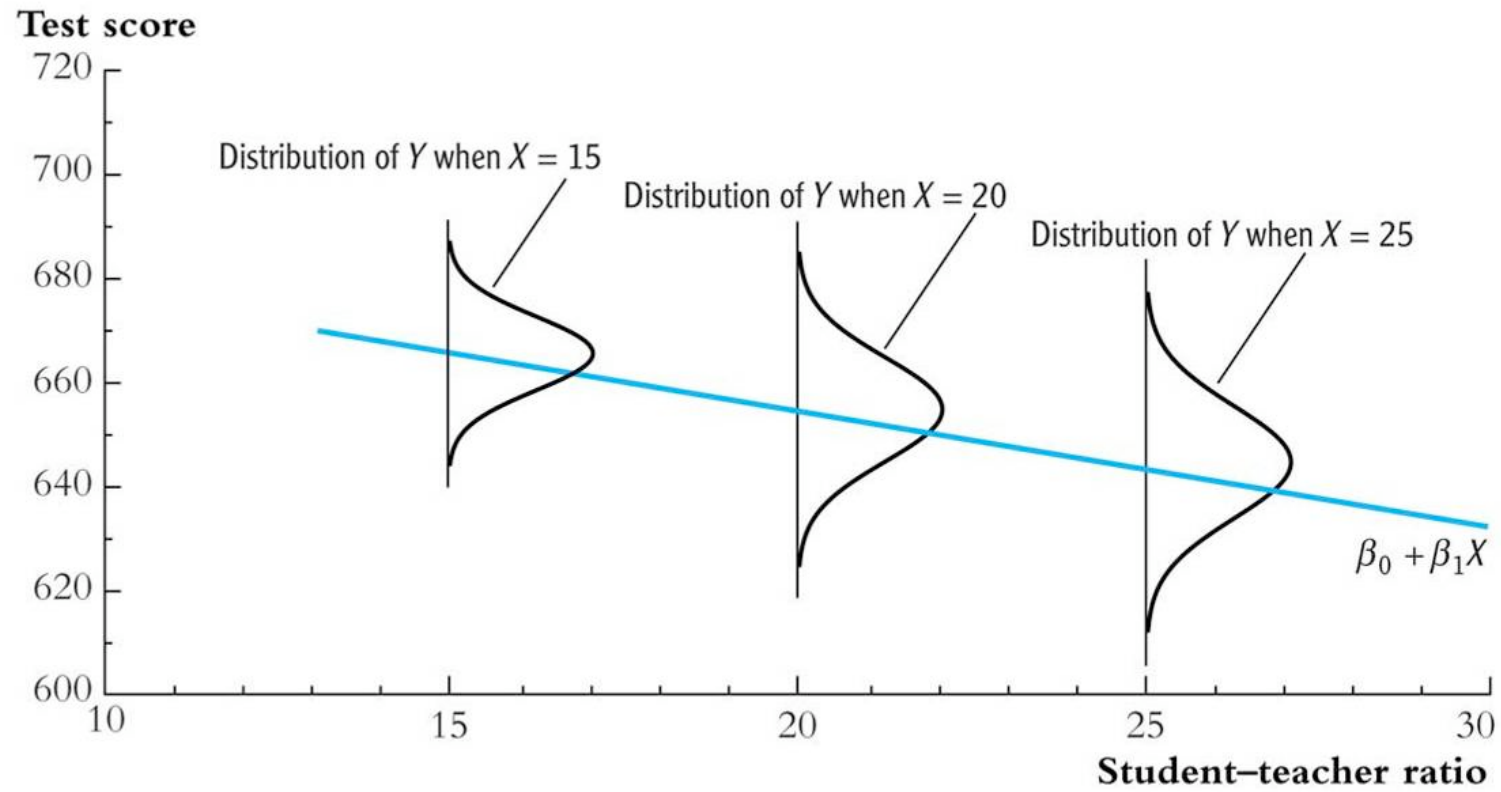
- È uno stimatore della deviazione standard dell'errore di regressione  $u_i$

$$SER = s_{\hat{u}}, \text{ dove } s_{\hat{u}}^2 = \frac{1}{n-2} \sum_{i=1}^n \hat{u}_i^2 = \frac{SSR}{n-2}$$

- Il denominatore è  $n-2$  gradi di libertà in quanto stimiamo i due coefficienti della regressione

# Omoschedasticità ed Eteroschedasticità

- L'errore di regressione  $u_i$  data  $X_i$  è costante per  $i=1, \dots, n$  e non dipende da  $X_i$ . Altrimenti l'errore è eteroschedasticità.
- Un [grafico](#) precedente mostra il caso in cui le distribuzioni condizionate dei punteggi dei test hanno la stessa dispersione. Cioè la varianza delle distribuzioni è la stessa per i diversi valori di  $x$ .
- La figura successiva mostra il caso in cui l'errore è eteroschedasticità.



- Nel caso in cui l'errore è omoschedastico e sono valide le assunzioni dei minimi quadrati, gli stimatori OLS sono i migliori stimatori lineari non distorti (BLUE). Questo è il teorema di Gauss-Markov
- In caso di eteroschedasticità gli stimatori OLS rimangono non distorti ed asintoticamente normali (consistenti), ma non sono più efficienti.
- Rimedio: usare errori standard robusti all'eteroschedasticità

# Esempio di regressione

```
regress testscr str , robust
```

Regression with robust standard errors

```
Number of obs =      420
F( 1, 418) =    19.26
Prob > F      =    0.0000
R-squared     =    0.0512
Root MSE     =    18.581
```

| -----   |  |           |           |       |       |                      |
|---------|--|-----------|-----------|-------|-------|----------------------|
|         |  | Robust    |           |       |       |                      |
| testscr |  | Coef.     | Std. Err. | t     | P> t  | [95% Conf. Interval] |
| -----   |  |           |           |       |       |                      |
| str     |  | -2.279808 | .5194892  | -4.39 | 0.000 | -3.300945 -1.258671  |
| _cons   |  | 698.933   | 10.36436  | 67.44 | 0.000 | 678.5602 719.3057    |
| -----   |  |           |           |       |       |                      |

- If you use the “**, robust**” option, STATA computes heteroskedasticity -robust standard errors
- Otherwise, STATA computes homoskedasticity -only standard errors